Ontology mapping with auxiliary resources

Summary

The availability of data plays an ever increasing role in our society. Businesses commonly store information about their customers, transactions and products in large information systems. This allows them to analyse their data to gain more knowledge, such as trends and predictions, in order to improve their business strategy. Furthermore, the core strategy of a business can be built on enabling the user to easily access a certain type of data. Such services play an increasing role in common every-day life. For example, services such as *Google* and *Wikipedia* are widely used to find general information, whereas services such as *Amazon*, *bol.com* and *Yelp* are used to find information and reviews about products. Some of these site also allow the user to purchase the queried products on the same site. To be able to interpret stored data, it is necessary that the data is structured and annotated with meta-information, such that for each data entry it is possible to determine its meaning and relation to other data entries. For example, a data entry '555-12345' has very little use if it is not known that it represents a telephone number and who the owner of the number is. An information system specifies this type of meanings and their structure using an ontology. An ontology specifies the types of objects, referred to as concepts, about which one intends to store information, what kind of data is stored for each concept and how the concepts are related to each other.

A common problem faced by businesses is the desire to be able to exchange information between different systems. An example scenario would be *Company A* deciding to acquire *Company B*. To continue the operations of *Company B*, *Company A* would need to transfer all the data of the information system of *Company B* into its own information system. Here, it can occur that the data in the information systems of both companies is modelled using different ontologies. This can stem from the companies having different requirements for their systems or having followed separate design principles in the creation of their ontologies. In this case, it is not possible to simply transfer data between the systems since these are incompatible.

A possible solution for enabling the exchange of information between systems utilizing different ontologies is the process of *ontology mapping*. Ontology mapping aims to identify all pairs of concepts between two ontologies which are used to model the same type of information. A full list of correspondences between two ontologies is known as an alignment or mapping. Based on such a mapping, it is possible to create a transfer function such that every data entry part of one ontology can be re-formulated such that it conforms to the specification of the other ontology. This allows for the transfer of data between two information systems despite the systems

using different ontology structures.

Mapping ontologies is a labour intensive task. To create a mapping, a domain expert has to manually define and verify every correspondence. This approach is infeasible when having to map large ontologies encompassing thousands of concepts. Hence, automatic approaches to ontology mapping are required in order to solver interoperability problems in the corporate domain. A different domain of application is the *Semantic Web*. This domain envisions the next step in the evolution of the world-wide-web, where all available information is machine readable and semantically structured. This semantic structure is also specified using an ontology and allows machines to gather semantic information from the web. However, in order to retrieve semantic information autonomously, a machine needs to be capable to also autonomously match ontologies. This is necessary such that the machine can query sources which represent their information using a different semantic structure.

Ontology mapping has been an active field of research in the past decade. Here, matching systems typically utilize a combination of techniques to determine the similarity between concepts. From these computations, highly similar concepts are extracted which then form the alignment between the given ontologies. In some situations, it is possible that an extra resource of information is available that can be exploited to aid the matching process. An example of such extra information are lexical resources, for instance *Wikipedia*. A lexical resource allows a system to look up word definitions, identify synonyms and look up information of related concepts. A different example resource are partial alignments. A partial alignment is an incomplete mapping stemming from an earlier matching effort. It can be the result of a domain expert attempting to create a mapping, but being unable to finish it due to time constraints. A core challenge within the field of ontology mapping thus is to devise techniques which can use these resources for the purpose of creating a complete mapping. This has led us to the following problem statement:

> *How can we improve ontology mapping systems by exploiting auxiliary information?*

To tackle this problem statement, we formulated four research questions upon which we based our research:

1. *How can lexical sense definitions be accurately linked to ontology concepts?*
2. *How can we exploit partial alignments in order to derive concept correspondences?*
3. *How can we evaluate whether partial alignment correspondences are reliable?*
4. *To what extent can partial alignments be used in order to bridge a large terminological gap between ontologies?*

In Chapter 1 we introduce the reader to the field of ontology mapping. Here, we introduce the problems that arise when attempting to transfer data between knowledge systems with different ontologies. Further, we present a series of real-world domains which can benefit from the research of this thesis, such as information

integration, web-service composition and agent communication. We also present a brief overview of the core research challenges of the field of ontology mapping. In the final section of the chapter, we introduce and discuss the problem statement and research questions which guide this thesis.

Chapter 2 provides important background information to the reader. We formally introduce the problem of ontology matching. Further, we detail and illustrate common techniques that are applied for the purpose of ontology alignment evaluation. Lastly, we introduce a series of datasets which can be used in order to evaluate ontology matching systems.

Applicable techniques to ontology matching are introduced in Chapter 3. Here, we introduce the reader to contemporary ontology matching systems and their underlying architectures. We introduce the three core tasks that a matching systems has to perform, being similarity computation, similarity combination and correspondence extraction, and provide an overview of techniques which are applicable for these respective tasks. Additionally, we provide a brief survey of existing ontology matching systems with the focus on systems utilizing auxiliary resources.

Chapter 4 answers the first research question. Here, the core problem concerns the linking of correct lexical definitions to the modelled concepts of an ontology, referred to as concept disambiguation. An example of such an action is determining that the concept name 'Plane' refers to the type of mathematical surfaces instead of the type of airborne vehicles. Techniques utilizing lexical resources rely on these links to determine concept similarities using various techniques. We tackle this research question by proposing an information retrieval based method for associating ontology concepts with lexical senses. Using this method, we define a framework for the filtering of concept senses based on sense-similarity scores and a given filtering policy. We evaluate four filtering policies which filter senses if their similarity scores resulted in an unsatisfactory value. The filtering policies are evaluated using several lexical similarities in order to investigate the general effects of the filtering policies on the lexical similarities. Our evaluation revealed that the application of our disambiguation approach improved the performances of all lexical metrics. Additionally, we investigated the effect of weighting the terms of the sense annotations and concept annotations. This evaluation revealed that weighting terms according to respective origins within the ontologies or lexical resource resulted in a superior performance compared to weighting the document terms using the widely used TF-IDF approach from the field of information retrieval.

The research question tackled in Chapter 5 concerns the exploitation of partial alignments. The core problem here is that a matching system is given an incomplete alignment, referred to as partial alignment, and has to compute the remaining correspondences to create a full mapping. For this purpose, one has to create mapping techniques which utilize the information of the individual partial alignment correspondences, referred to as *anchors*, in order to improve the mapping quality. To answer this question, we propose a method which compares concepts by measuring their similarities with the provided anchor concepts. For each concept, its measurements are compiled into an *anchor-profile*. Two concepts are then considered similar if their anchor-profiles are similar, i.e. they exhibit comparable degrees of similarity towards the anchor concepts. The evaluation revealed that the applica-

tion of our approach can result into performances similar to top matching systems in the field. However, we observe that the performance depends on the existence of appropriate meta-information which is used to compare concepts with anchors. From this, we conclude that a combination of similarity metrics, such that all types of meta-information are exploited, should be used to ensure a high performance for all types of matching problems. Lastly, we systematically investigate the effect of the partial alignment size and correctness on the quality of the produced alignments. We observe that both size and correctness have a positive influence on the alignment quality. We observe that decreasing the degree of correctness has a more significant impact on the alignment quality than decreasing the size. From this we conclude that matching systems exploiting partial alignments need to take measures to ensure the correctness of a given partial alignment from an unknown source.

Chapter 6 addresses research question 3, which presents the problem of ensuring the correctness of a partial alignment. Some techniques exploit partial alignments for the purpose of ontology mapping. An example of such a technique is the approach presented in Chapter 5. In order for these techniques to function correctly, it is necessary that the given partial alignment contains as few errors as possible. To evaluate the correctness of a given partial alignment, we propose a method utilizing feature evaluation techniques from the field of machine learning. To apply such techniques, one must first define a feature space. A feature space is a core mathematical concept describing a space that is spanned by inspecting n different variables. For example, taking the variables 'height', 'width' and 'depth' would span a 3-dimensional feature space. Plotting the respective values for each feature of different objects would thus allow us to inspect the differences between the objects with regard of their physical location and perform analytical tasks based on this data. In the field of machine learning, a feature space is not restricted by the amount or types of features. Therefore, one can span a feature space using any amount of features modelling quantities such as position, size, age, cost, type or duration. A core task in the field of machine learning is classification, where one must designate a class to an object for which the values for each feature are known. An example of such a task is determining whether a person is a reliable debtor given his or her income, employment type, age and marital status. A classification system does this by first analysing a series of objects for which the class values are already known. Feature evaluation techniques help the designer of a specific classification system to determine the quality of a feature with respect to the classification task by analysing the pre-classified objects. For our approach, we utilize this work-flow in order to design an evaluation system for a series of anchors. We span a feature space where every feature represents the result of a consistency evaluation between a specific anchor and a given correspondence. Using a selection of feature evaluation techniques, we then measure the quality of each feature and therefore the quality of its corresponding anchor. To generate the consistency measurements, we define a metric requiring a base similarity, for which we evaluate three types of similarities: a syntactical, a profile and a lexical similarity. For each type of similarity, we evaluate our approach against a baseline ranking, which is created by directly applying the same similarity on each anchor. Our evaluation revealed that our approach was able to produce better anchor evaluations for each type of similarity metric than

the corresponding baseline. For the syntactic and lexical similarities we observed significant improvements.

The research presented in Chapter 7 tackles the fourth research question. This chapter focuses on a specific kind of matching problems, being ontologies which have very little terminology in common. Many matching techniques rely on the presence of shared or very similar terminology in order to decide whether two concepts should be matched. These techniques fail to perform adequately if the given ontologies use different terminology to model the same concepts. Existing techniques circumvent this problem by adding new terminology to the concept definitions. The new terms can be acquired by searching a lexical resource such as *WordNet*, *Wikipedia* or *Google*. However, if an appropriate source of new terminology is not available then it becomes significantly harder to match these ontologies. We investigate a possible alternative by proposing a method exploiting a given partial alignment. Our approach is built upon an existing profile similarity. This type of similarity exploits semantic relations in order to gather context information which is useful for matching. Our extension allows it to exploit the semantic relations that are specified in the partial alignment as well. The evaluation reveals that our approach can compute correspondences of similar quality exploiting only partial alignments as existing frameworks using appropriate lexical resources. Furthermore, we establish that a higher performance is achievable if both lexical resources and partial alignments are exploited by a mapping system.

Chapter 8 provides the conclusions of this thesis and discusses possibilities of future research. Taking the answers to the research questions into account, we conclude that there are a multitude of ways in which auxiliary resources can be exploited in order to aide ontology matching systems. First, lexical resources can be effectively exploited when applying a virtual document-based disambiguation policy. Second, through the creation of anchor-profiles it is possible to exploit partial alignments to derive similarity scores between concepts. Third, by using a feature-evaluation approach one can evaluate anchors to ensure approaches utilizing partial alignments perform as expected. Fourth, by extending profile similarities, such that these also exploit anchors, one can match ontologies with little to no terminological overlap.

When discussing future research, we identify several key areas which should be investigated to improve the applicability of the presented work. First, research efforts should be directed into the robustness of the approaches. For example, the disambiguation approach of Chapter 4 relies on the presence of terminological information to be able to identify senses. If this information is sparse or lacking all together, then the effectiveness of this approach can be affected. A solution could be the combination of multiple disambiguation approaches. Based on the available meta-information, a decision system could determine which approach is best suited for each ontology. A different area for future research would be the generation of reliable partial alignments. If a partial alignment does not exist for a given matching problem, then generating one during run-time would enable a matching system to use techniques which require the existence of such alignments. This would allow for the presented techniques of this thesis to be applicable to a wider group matching problems.

Samenvatting

De beschikbaarheid van data speelt een steeds belangrijkere rol in onze samenleving. Bedrijven gebruiken vaak informatiesystemen om informatie op te slaan over hun klanten, transacties en producten. Daardoor is het mogelijk om deze data te analyseren en om zo meer kennis te (her)gebruiken door bijvoorbeeld voorspellingen te doen aan de hand van trends, zodanig dat bedrijven hun handelsstrategieën kunnen verbeteren. Sterker nog, een bedrijf kan zich richten op het toegankelijk maken van databronnen voor consumenten. Dit soort diensten speelt een steeds grotere rol in het alledaagse leven. Diensten zoals *Google* en *Wikipedia* worden doorgaans gebruikt om algemene informatie te vinden. Gespecialiseerde diensten zoals *Amazon*, *bol.com* en *Yelp* worden gebruikt om informatie en beoordelingen van producten te vinden en om deze producten zelfs te kopen. Om opgeslagen data te kunnen interpreteren is het noodzakelijk dat deze data een structuur heeft en geannoteerd is met meta-informatie. Dit maakt het mogelijk om de betekenis van ieder datapunt, en zijn relatie met andere datapunten, te bepalen. Bijvoorbeeld, het datapunt '555-12345' is van weinig nut als het niet bekend is dat dit een telefoonnummer representeert en wie de eigenaar van dit nummer is. Een informatiesysteem beschrijft de betekenissen en structuur van de opgeslagen data met behulp van een zogenaamde ontologie. Deze ontologie specificeert een aantal typen, ookwel concepten genoemd, en hoe deze concepten aan elkaar zijn gerelateerd.

Bedrijven staan vaak voor het probleem dat ze informatie willen uitwisselen tussen verschillende systemen. Veronderstel bijvoorbeeld dat *Bedrijf A* beslist om *Bedrijf B* over te nemen. Om de bedrijfsvoering van *Bedrijf B* voort te kunnen zetten moet *Bedrijf A* alle informatie uit het systeem van *Bedrijf B* in zijn eigen systeem overzetten. Hier kan het gebeuren dat de data van *Bedrijf B* met een andere ontologie is gemodelleerd dan de data van *Bedrijf A*. De oorzaak hiervan kan zijn dat de twee bedrijven verschillende eisen hebben voor hun systemen of verschillende ontwerpprincipes hebben gehanteerd bij het definiëren van de ontologieën. In dit soort gevallen is het door de incompatibiliteit van de systemen niet zomaar mogelijk om data tussen de twee systemen uit te wisselen.

Een mogelijke oplossing om het overzetten van data tussen twee systemen, welke verschillende ontologieën gebruiken, mogelijk te maken is het zogenoemde *ontologie-mapping* proces. Het doel van ontologie-mapping is het identificeren van alle conceptparen welke gebruikt kunnen worden om dezelfde soort data te modelleren. Een volledige lijst van correspondenties tussen twee ontologieën wordt opgeslagen in een zogenaamde *alignment* of *mapping*. Met behulp van deze mapping is het mogelijk

om data uit één ontologie te herschrijven zodat deze conform is aan de specificaties van de andere ontologie. Hierdoor wordt het mogelijk om data tussen twee systemen uit te wisselen, ondanks het feit dat de twee systemen verschillende ontologieën gebruiken. Het maken van zo'n mapping vergt veel werk. Om een mapping te maken moet een domeinexpert handmatig alle correspondenties definiëren en controleren. Deze aanpak is niet haalbaar als men een mapping tussen twee grote ontologieën moet maken, waarbij iedere ontologie duizenden concepten modelleert. Het is dus nodig om het proces van ontologie-mapping te automatiseren. Een ander applicatiedomein is het zogenaamde *Semantic Web*. Dit domein stelt de volgende stap in de evolutie van het world-wide-web voor, waar alle beschikbare informatie door een machine leesbaar en semantisch gestructureerd is. Deze semantische structuur is ook gedefinieerd door middels een ontologie, zodanig dat het machines mogelijk is om semantische informatie uit het web te verzamelen. Om semantische informatie onafhankelijk te verzamelen, moet het voor een machine mogelijk zijn om verschillende ontologieën automatisch te mappen. Met behulp van een mapping is het voor een machine mogelijk om informatie te verzamelen welke in een verschillende semantische structuur is gemodelleerd.

Sinds het afgelopen decennium is ontologie-mapping een actief onderzoeksveld. Er zijn gespecialiseerde mapping-systemen ontwikkeld die gebruik maken van een combinatie van technieken om de overeenkomsten tussen concepten te bepalen. Met behulp van deze systemen worden overeenkomende concepten geëxtraheerd, welke vervolgens de alignment tussen de twee ontologieën vormen. In sommige gevallen is het mogelijk dat er extra informatie beschikbaar is, welke gebruikt kan worden om het mapping-proces te verbeteren. Een voorbeeld van dit soort extra informatie is *Wikipedia*. Een lexicale bron zoals Wikipedia maakt het mogelijk om definities van woorden te raadplegen, synonieme woorden te identificeren en informatie van gerelateerde concepten te raadplegen. Een ander voorbeeld van een extra informatiebron is een partiële mapping. Een partiële mapping is een onvolledige mapping welke het resultaat is van een eerdere poging om een mapping tussen de ontologieën te creëren. Deze mapping is onvolledig omdat bijvoorbeeld een domainexpert niet in staat was deze te voltooien wegens tijdgebrek. Een belangrijke uitdaging in het veld van ontologie-mapping is dus het creëren van technieken welke van dit soort informatiebronnen gebruik maken om een mapping te genereren. Dit heeft ons naar de volgende probleemstelling geleid:

> *Hoe kunnen we ontologie-mapping systemen verbeteren door gebruik te maken van externe informatiebronnen?*

Om deze probleemstelling aan te pakken hebben we vier onderzoeksvragen geformuleerd welke dit onderzoek gestuurd hebben:

1. *Hoe kan men nauwkeurig lexicale betekenissen aan ontologieconcepten koppelen?*

2. *Hoe kan men partiële mappings gebruiken om de overeenkomsten tussen concepten te bepalen?*

3. *Hoe kan men beoordelen of correspondenties afkomstig uit partiële mappings betrouwbaar zijn?*

4. *In hoeverre is het mogelijk om parti´ele overeenkomsten te gebruiken om mappings tussen ontologieŁen te genereren die weinig overeenkomstige terminologieën gebruiken?*

In hoofdstuk 1 introduceren we het onderzoeksveld van ontologie-mapping. We introduceren de problemen die kunnen ontstaan als men data tussen informatiesystemen wil uitwisselen. Verder introduceren wij een reeks van reële domeinen waar het gepresenteerde werk van toepassing is, zoals bijvoorbeeld informatie integratie, webdienst-compositie en agentcommunicatie. Wij presenteren ook een kort overzicht van de belangrijkste onderzoeksproblemen met betrekking tot ontologie-mapping. In de laatste sectie van dit hoofdstuk introduceren en bespreken wij de probleemstelling en de onderzoeksvragen van dit proefschrift.

Hoofdstuk 2 maakt de lezer bekend met belangrijke achtergrondinformatie. Hier introduceren wij formeel het probleem van ontologie-mapping. Verder detailleren en illustreren we de meest gebruikte technieken waarmee mappings geëvalueerd kunnen worden. Tot slot introduceren wij een aantal datasets die gebruikt kunnen worden om een ontologie-mapping systeem te evalueren.

Technieken die beschikbaar zijn voor het maken vaan een ontologie-mapping worden in hoofdstuk 3 geïntroduceerd. Hier maken wij de lezer bekend met de opbouw van huidige mapping-systemen en de meest gebruikte technieken. Wij introduceren hier de drie kerntaken welke een mapping-systeem moet uitvoeren, namelijk de overeenkomstberekening, overeenkomstcombinatie en correspondentie-extractie. Voor iedere kerntaak geven wij een overzicht van technieken die voor de gegeven taak van toepassing zijn. Vervolgens geven wij een overzicht van huidige mapping-systemen met een focus op systemen die gebruik maken van externe informatiebronnen.

In hoofdstuk 4 beantwoorden wij de eerste onderzoeksvraag. Het kernprobleem hier betreft het nauwkeurig koppelen van lexicale definities aan de gemodelleerde concepten uit de ontologie. Dit proces staat bekend onder de naam *disambiguatie*. Een voorbeeld van dit proces is het vaststellen dat het concept 'Bank' naar het financiële instituut refereert en niet naar het meubelstuk. Technieken die van lexicale informatiebronnen gebruik maken hebben deze koppelingen nodig om de overeenkomsten tussen concepten te bepalen met behulp van bepaalde algoritmen. Wij pakken deze onderzoeksvraag aan door het introduceren van een op Information-Retrieval-gebaseerde techniek waarmee ontologieconcepten aan lexicale betekenissen gekoppeld kunnen worden. Met behulp van deze techniek zetten wij een disambiguatie-kader op waarmee lexicale bedoelingen gefilterd worden afhankelijk van hun overeenkomstwaarden en een filterstrategie. Wij evalueren vier verschillende filterstrategieën die een lexicale betekenis filteren als ze de bijbehorende overeenkomstwaarde onvoldoende vinden. De filterstrategieën worden geëvalueerd met behulp van drie verschillende lexicale overeenskomstmetrieken. Onze evaluatie heeft laten zien dat het toepassen van onze disambiguatieaanpak de prestaties van alle drie overeenkomstmaten heeft verbeterd. Verder hebben wij het effect van het verzwaren van de termgewichten van de concept-annotaties en betekenis-annotaties

onderzocht. Deze evaluatie heeft laten zien dat het verzwaren van termgewichten afhankelijk van hun oorsprong in de ontologie of lexicale informatiebron een groter positief effect heeft op de prestatie dan het toepassen van het veelgebruikte TF-IDF aanpak, afkomstig uit het veld van Information-Retrieval.

Het onderzoek in hoofdstuk 5 behandelt onderzoeksvraag 2 in relatie tot het gebruiken van partiële mappings. Het kernprobleem hier is dat het mapping-systeem toegang tot een onvolledige mapping heeft, ook wel een partiëel-mapping genoemd, en dus de onbekende correspondenties moet bepalen om een volledige mapping te creëren. Hier is het doel om de individuele correspondenties van de partiële mapping, ook wel *ankers* genoemd, te gebruiken om de kwaliteit van de berekende mappings te verbeteren. Om deze vraag te beantwoorden stellen wij een methode voor die is gebaseerd op het vergelijken van concepten door het meten van de overeenkomsten tussen een concept en de gegeven ankers. Aan ieder concept worden de overeenkomstwaarden in een zogenoemd *ankerprofiel* samengevoegd. Twee concepten worden als overeenkomend beschouwd als hun ankerprofielen overeenkomen, d.w.z. dat zij vergelijkbare overeenkomsten hebben met de ankerconcepten. In onze evaluatie hebben wij kunnen vaststellen dat onze aanpak in staat is om prestaties te leveren die vergelijkbaar zijn met de top mapping-systemen in het gebied. Onze aanpak is echter wel afhankelijk van het bestaan van geschikte meta-informatie waarmee concepten met ankers worden vergeleken. Hieruit concluderen wij dat alle soorten van meta-informatie geraadpleegd moeten worden door een combinatie van overeenkomstmaten toe te passen om zeker te zijn dat deze techniek voor alle soorten problemen geschikt is. Tot slot voeren wij een systematisch onderzoek uit om vast te stellen hoe groot de invloed is van de grootte en de correctheid van de partiële mapping op de kwaliteit van de berekende mapping. Hier stellen wij vast dat zowel de grootte als ook de correctheid invloed hebben op de mappingkwaliteit. Verder stellen wij vast dat een vermindering van de correctheid van de partiële mapping een sterkere invloed heeft dan een vermindering van de grootte. Hieruit concluderen wij dat mapping-systemen die van partiële mappings gebruik maken maatregelen moeten nemen om ervoor te zorgen dat partiële mappings uit onbekende bronnen correct zijn.

Onderzoeksvraag 3 is het hoofdthema van hoofdstuk 6. De kernvraag hier is het zeker stellen van de correctheid van een gegeven partiële mapping. Sommige technieken genereren een mapping met behulp van een partiële mapping. Een voorbeeld van zo'n techniek is te zien in hoofdstuk 5. Het is voor deze technieken nodig dat de gegeven partiële mapping zo min mogelijk fouten bevat om er voor te zorgen dat deze technieken adequaat presteren. Om de correctheid van een partiële mapping te evalueren stellen wij een techniek voor die gebruik maakt van feature-evaluatietechnieken, afkomstig uit het veld van machine-learning. Om van een feature-evaluatietechniek gebruik te maken moet men eerst een feature-ruimte definiëren. Een feature-ruimte is een kernconcept uit de wiskunde die een ruimte beschrijft die wordt opgespannen door n verschillende features. Bijvoorbeeld, door gebruik te maken van de features 'hoogte', 'breedte' en 'diepte' kan men een 3-dimensionale feature-ruimte opspannen. Door de bijbehorende waardes van de verschillende objecten in deze ruimte te plotten kan men de verhouding van de objecten zien in verband met hun fysieke locatie, en met behulp van deze data verschillende

analyses uitvoeren. In het veld van machine-learning zijn er geen beperkingen wat betreft het soort of aantal van gebruikte features. Het is dus ook mogelijk om features te gebruiken om maten zoals positie, grootte, leeftijd, kosten, type of duur weer te geven. Een kerntaak in het veld van machine-learning is classificatie, waar men een categorie aan een object moet toekennen voor welke de waarde van ieder feature bekend is. Een voorbeeld van zo'n taak is het bepalen of een persoon een betrouwbare debiteur is, afhankelijk van zijn inkomen, soort aanstelling, leeftijd en gezinstoestand. Een classificatiesysteem doet dit door eerst een reeks objecten te analyseren van welke de bijbehorende categorieën al bekend zijn. Feature-evaluatietechnieken helpen de maker van een classificatiesysteem te bepalen hoe zeer een bepaalde feature van nut is m.b.t. de classificatietaak door middel van het analyseren van de al geclassificeerde objecten. Voor onze aanpak benutten wij deze werk-flow om een evaluatiesysteem voor een reeks ankers te creëren. Wij zetten een feature-ruimte op waar elke feature het resultaat van een consistentie-evaluatie tussen een specifiek anker en gegeven correspondentie representeert. Met behulp van feature-evaluatietechnieken evalueren wij de kwaliteit van de features, en dus ook de kwaliteit van de bijbehorende ankers. Om de consistentie-waarden te berekenen, definiëren wij een maat welke gebruik maakt van een basis overeenkomstmetriek. Wij evalueren drie soorten maten als basis overeenkomstmetriek: een syntactische, een profiel en een lexicale metriek. Voor ieder type maat evalueren wij onze aanpak ten opzichte van een basis evaluatie, welke gemaakt is door het directe toepassen van de maat op de ankers. Onze evaluatie heeft laten zien dat voor ieder soort maat onze aanpak betere evaluaties produceert dan de bijbehorende basis evaluatie. Voor de syntactische en lexicale maat hebben wij significante verbetering vast kunnen stellen.

Onderzoeksvraag 4 is het onderwerp van het onderzoek dat is beschreven in hoofdstuk 7. Het hoofdthema hier is een specifieke soort van mapping-taken, namelijk het mappen van ontologieën die verschillende terminologieën gebruiken om dezelfde concepten te modelleren. Veel technieken vereisen het bestaan van gelijke of soortgelijke termen om te bepalen of twee concepten met elkaar overeen komen of niet. Deze technieken presteren slecht als de gegeven ontologieën terminologieën gebruiken die doorgaans verschillend zijn. Bestaande technieken vermijden dit probleem door middel van het toevoegen van nieuwe terminologie aan de conceptdefinities. Nieuwe termen worden opgezocht door bijvoorbeeld het raadplegen van bronnen zoals *WordNet*, *Wikipedia* of *Google*. Als dit soort bronnen niet beschikbaar zijn, dan wordt het aanzienlijk moeilijker om een mapping tussen de twee ontologieën te maken. Wij onderzoeken een alternatief dat van een gegeven partiële mapping gebruik maakt. Onze aanpak is gebaseerd op bestaande profiel-overeenkomstmetrieken. Dit soort maten maakt gebruik van semantische relaties om belangrijke contextinformatie te verzamelen. Onze uitbreiding van de gebruikte maat maakt het mogelijk om ook van de semantische relaties in de partiële mapping gebruik te maken. Onze evaluatie heeft laten zien dat onze aanpak in staat is om met behulp van alleen een partiële mapping correspondenties te genereren die een vergelijkbare kwaliteit hebben als bestaande technieken die van lexicale informatiebronnen gebruik maken. Verder stellen wij vast dat een betere kwaliteit haalbaar is als een techniek gebruik maakt van zowel lexicale bronnen als ook partiële mappings.

Hoofdstuk 8 geeft de conclusies van dit proefschrift en bespreekt de mogelijkhe-

den voor verder onderzoek. Rekening houdend met de antwoorden op de gestelde onderzoeksvragen concluderen wij dat er diverse mogelijkheden bestaan om externe informatiebronnen te gebruiken voor het verbeteren van ontologie-mapping-systemen. Ten eerste kunnen lexicale informatiebronnen beter benut worden als eerst een disambiguatie methode toegepast wordt. Ten tweede, door het creëren van ankerprofielen is het mogelijk om overeenkomstwaarden tussen concepten te berekenen. Ten derde, door gebruik te maken van een op feature-selectie-gebaseerde techniek is het mogelijk om ankers te evalueren en er zo voor te zorgen dat technieken die van partiële mappings gebruik maken presteren zoals kan woorden verwacht. Ten vierde, door het uitbreiden van profiel-overeenkomstmetrieken, zodanig dat deze van gegeven ankers gebruik maken, is het mogelijk om ontologieën te mappen die weinig terminologie met elkaar gemeen hebben.

In de discussie over verder onderzoek stellen wij verschillende onderwerpen vast die onderzocht moeten worden om de toepasbaarheid van het gepresenteerde onderzoek te verbeteren. Een onderwerp voor verdergaand onderzoek is dat de robuustheid van de technieken onderzocht moet worden. Bijvoorbeeld, de prestatie van de disambiguatie-techniek van hoofdstuk 4 hangt af van de aanwezigheid van terminologische informatie om zo de correcte lexicale bedoelingen te identificeren. Bij ontologieën waar weinig tot helemaal geen terminologische informatie is gemodelleerd zou het mogelijk kunnen zijn dat de prestaties van de aanpak slechter uitvallen dan verwacht. Een mogelijke oplossing zou het combineren van verschillende disambiguatie-technieken kunnen zijn. Afhankelijk van de beschikbare metainformatie zou een beslissysteem kunnen bepalen welke disambiguatie-techniek het meest geschikt is voor elke ontologie. Een ander gebied voor verder onderzoek zou het generen van partiële mappings kunnen zijn. Voor problemen waar geen partiële mappping beschikbaar is, zou het genereren van een betrouwbare partiële mapping het mogelijk maken om technieken die gebruik maken van deze partiële mappings toe te passen. Dit zou het dus mogelijk maken om de gepresenteerde technieken op een grotere reeks problemen toe te passen.